11.3 Describing and Analyzing Data – Overview

Measures of Center

Mean (Average Value)

$$\boxed{\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}}$$

- Simple, arithmetic average of the data.
  - Sum all numbers and divide by the sample size (n).
- Same calculation for the population and sample mean (just different notation).
  - Sample mean = $\bar{x}$ (pronounced "x-bar")
  - Population mean = μ (Greek letter mu)
- Mean is NOT a resistant measure.
  - This means it is heavily affected by outliers.

Example 1 – Mean

Data: 1, 5, 2, 9, 3

$\bar{x} = \frac{1+5+2+9+3}{5} = 4$

new $\bar{x}$ (with 30) = 9.4

Median (Middle Value)

- The middle value in an ordered list.
- Median IS a resistant measure.
  - NOT affected by outliers.

Example 2 – Median

Case 1 – Odd n

7 Obs: 10, 5, 6, 1, 3, 9, 8

Sorted: ~~1~~, ~~3~~, ~~5~~, 6, ~~8~~, ~~9~~, ~~10~~

Med = 6

Case 2 – Even n

8 Obs: 10, 5, 6, 1, 3, 9, 8, 3

Sorted: ~~1~~, ~~3~~, ~~3~~, 5, 6, ~~8~~, ~~9~~, ~~10~~

Med = $\frac{5+6}{2} = 5.5$

Mode (Most Common Value)

- The most frequently occurring value(s).
  - Unimodal data has one mode. → Ex) Case 2, Mode = 3 (twice)
  - Bimodal data has 2 modes.
  - Multimodal data has more than 2 modes.
  - Can be no modes (every value is distinct). → Ex) Case 1, no mode
- This is the only measure of center that can be used with categorical data.
  - Ex) Most common favorite color (can't average this)

Measures of Spread (Dispersion)

Range

- $\boxed{\text{Range = Max - Min}}$
- Gives idea of the entire "range" of values, how much distance do they span in total.
- Ex) Case 2 above: Range = 10 − 1 = 9

1 10

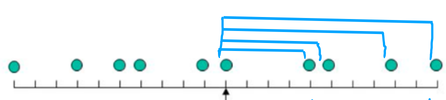
Standard Deviation

- Complex formula that measures the average distance that each data point is from the mean.

Sample Standard Deviation = S = $\sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$

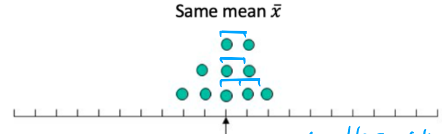
Population Standard Deviation = $\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$

(Greek letter sigma)



larger st dev

Same mean $\bar{x}$



smaller st dev

Using your Calculator!

Using TI-83/84 (and TI-30 XS MultiView / XIIS) to calculate mean, median, sample / population st dev.

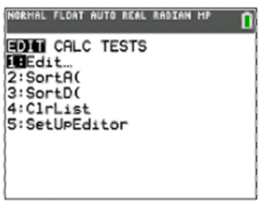
Steps for the TI-83/84	Steps for the TI-30XS MultiView	Steps for the TI-30 XIIS
1. Enter data: STAT → Edit → Enter data in L_1 *(Demo dataset: 10, 23, 4, 6, 9, 3, 15, 6)* 2. Calculate: STAT → CALC → 1-Var Stats a) List is L_1. b) Leave FreqList blank. c) Calculate!	1. Data → Enter data in L_1 2. 2^{nd} → stat → 1-Var Stats a) DATA: L1 b) FRQ: ONE c) CALC	1. 2^{nd} → STAT → 1-VAR (Enter) 2. DATA X_1 = # (scroll down) FRQ = 1 (for ALL Xs, scroll down) X_2 = # (scroll down) … 3. STATVAR (scroll across) 4. To exit this menu: 2^{nd} → EXIT STAT → Y

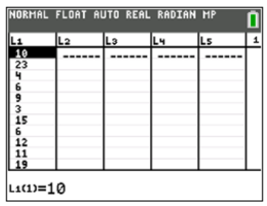
(does not give median)

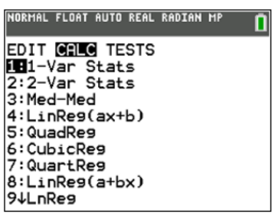
Inputs

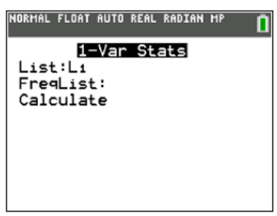
Data here:

10, 23, 4, 6, 9, 3, 15, 6, 12, 11, 19, 10, 6, 8, 15

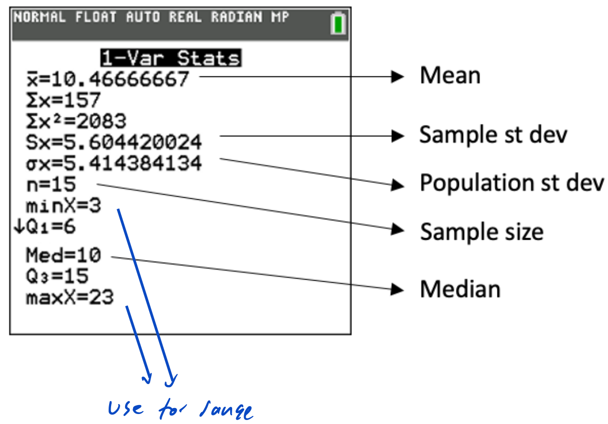








Results



Example 3

Find the mean, median, mode and sample standard deviation of the following dataset.

- Data (7 obs): 35, 70, 31, 37, 65, 38, 38

Results

Mean $\bar{x}$ = 44.86

Sample st dev S_x = 15.77

Pop st dev σ_x = 14.55

Med = 38

Other Considerations

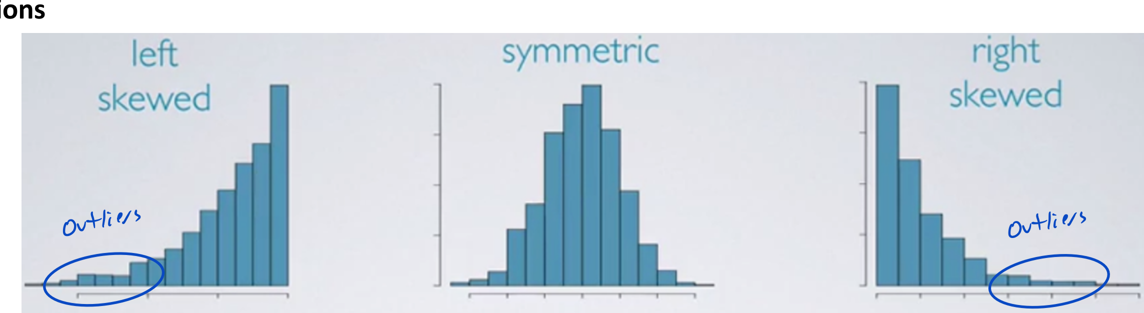
Outliers

ex) 35, 37, 38, 38, 65, 70

outliers

- Data values that are extreme when compared to the rest of the data.
- Can significantly impact measures of center and spread.

Types of Distributions



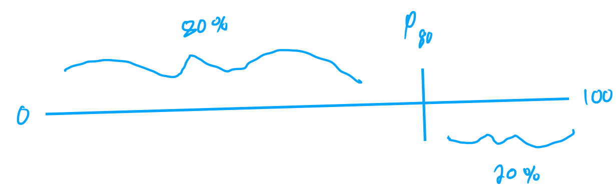
Best measure of center: Median mean Median

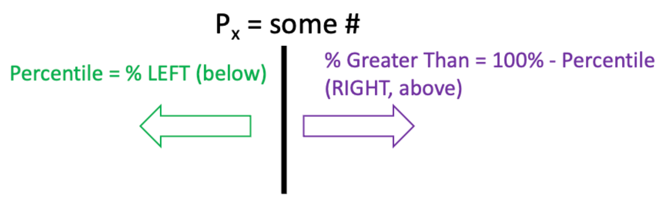
Measures of Relative Position

Percentiles

- A **percentile** tells you the percent of observations/individuals you are higher than.
  - Interpreting example: You are told you scored in the 90th percentile on GRE. This means you have a score that is higher than 90% of all others that took the test.
  - Range from 0th to 100th percentile.
  - There is complement aspect to percentiles as well; for example, if you are the 80th percentile, there is __20 %__ greater than you!

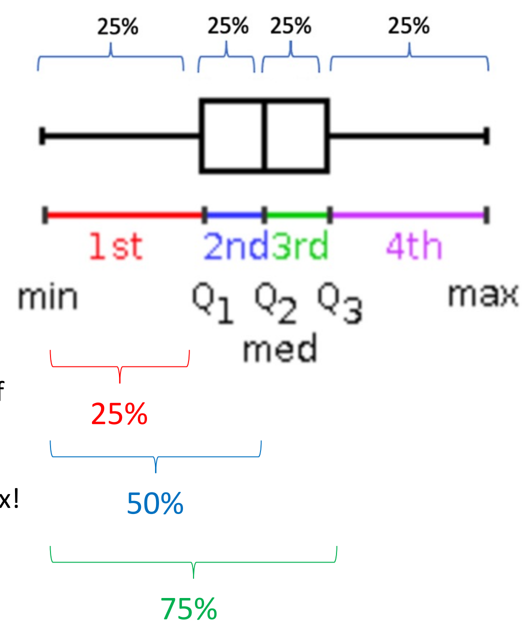
- Best way to remember!





- Notation: X^{th} Percentile = P_x

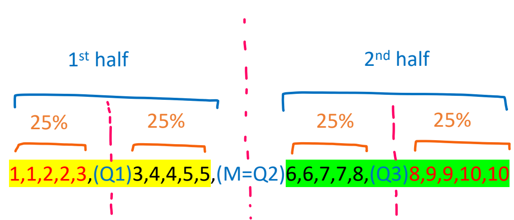
5-Number Summaries and Boxplots

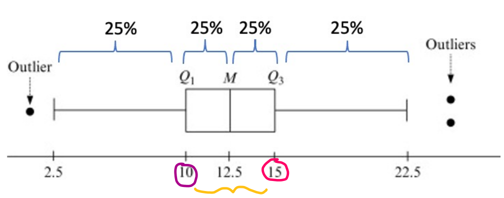


- **Quartiles** are specific percentiles.
  - Q_1 is the 25th Percentile.
  - Q_3 is the 75th Percentile.
  - Q_2 is the 50th Percentile = Median.

- **Inner Quartile Range (IQR)**
  - Another measure of variation, less informative than the standard deviation.
  - Uses quartiles to measure how far data is spread out around the median. Specifically, it measures the range of the middle 50% of the data
    - $\boxed{IQR = Q_3 - Q_1}$
  - Visualized very well in boxplots! It is the length of the box!

- **5-number summary**
  - $\boxed{\text{Min, } Q_1\text{, Med, } Q_3\text{, Max}}$ → Points of a boxplot

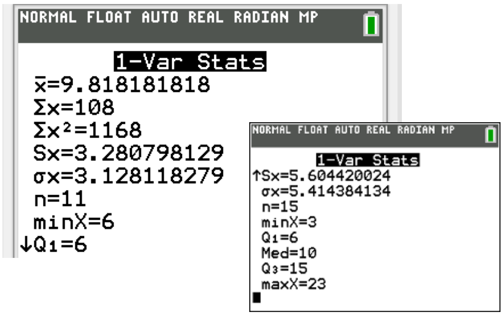




IQR = Q3 – Q1 = 15 – 10 = 5

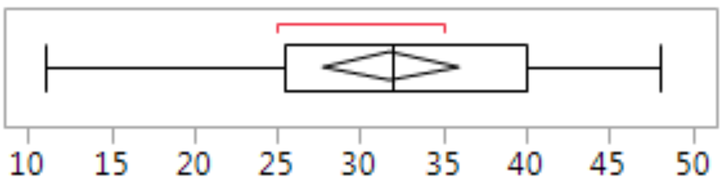
Example 4

a) Using this output from a 1-Var Stat, what is the IQR?



$IQR = Q_3 - Q_1$
$= 15 - 6$
$= 9$

b) Find the IQR from this boxplot.



$IQR = 40 - 25 = 15$

Example 5

a) Calculate the 5-number summary of the following dataset (20 numbers):

38, 33, 5, 5, 47, 29, 24, 42, 3, 18, 30, 46, 25, 44, 40, 42, 39, 44, 29, 13

By hand → 3, 4, 5, 5, 13, 18, 24, 25, 29, 29, 30, 33, 38, 39, 40, 42, 42, 44, 46, 47

min = 3

$Q_1 = \frac{13+18}{2} = 15.5$

Med $= \frac{29+30}{2} = 29.5$

$Q_3 = \frac{40+42}{2} = 41$

max = 47

Calculator → 1-Var stats (L_1 = #s) →
min = 3
Q_1 = 15.5
Med = 29.5
Q_3 = 41
Max = 47

b) Draw a boxplot [redacted] based on the 5-number summary from (a).